09 | Nvidia 与AI 芯片：超越摩尔定律

2021-06-07 邵巍

《说透芯片》

课程介绍 >

讲述：邵巍

时长 17:50　大小 16.34M

你好，我是邵巍。

我曾经在一个财经新闻中看到过一段话：“以纳米尺寸工艺制造的芯片，是人类自工业革命至今能够量产的最精巧、最神奇的造物。目前高性能芯片的运算能力早已远超人类，且正在源源不断地帮人类把人工智能（AI）这种过往只存在于科幻小说和科学家们头脑中的概念转化为现实。”

这一段话，我知道搞财经投资的朋友们有自己的目的，但是他们确实把半导体的行业之光写得非常文艺也非常到位。以纳米尺寸工艺制造芯片，那是我上一讲分享的台积电的功绩；而把 AI、科学家头脑中的概念转化为现实的，这就是我今天要好好讲讲的英伟达。

另外要提一下，财经投资的朋友们可不是只动嘴巴夸夸英伟达，他们真的下场投资。如果论营收，看 2021 年 4 月中旬的数字，Intel 是英伟达的近 5 倍，但是如果论市值，英伟达是 Intel 的 1.5 倍。为什么？凭什么？营收算是行业内部给予一个公司的评价，真金白银的买产品；而市值，就是大众对于公司的社会价值的肯定，那么大众肯定的是什么？是跟我一样，看到的是超行业、超摩尔定律的速度？还是大众对于 AI 这个大浪潮更有想象力？

今天，我就给就你讲一讲英伟达这家公司，是如何以 GPU 开局，并逐步在 AI 领域占据绝对优势的。

GPU 开局

让我先从英伟达的主要产品说起。看看下面这张英伟达财报中的表格，2020 年，英伟达收购成功 Mellanox，表格上“Compute&Networking”那一行的收入就来源于此，如果不算这部分，英伟达就是一个 GPU 的公司。

Revenue by Reportable Segments			
($ in millions)	FY21	FY20	Y/Y
Graphics	$9,834	$7,639	Up 29%
Compute & Networking	6,841	3,279	Up 109%
Total	$16,675	$10,918	Up 53%

我在基础知识篇，给你讲过了什么是 GPU，通俗一点理解，GPU 就是加速绘图的处理器。

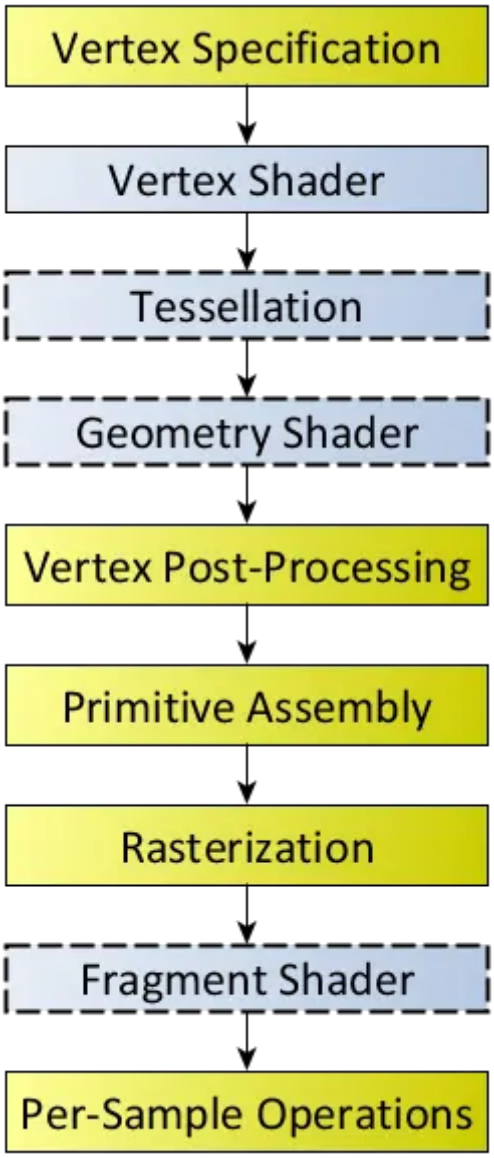
关于绘图的任务，我可以给你讲讲大概的原理。现在的显示器分辨率是 1080p，就是说这个显示器上的所有的图形，都是由水平方向 1920 行，垂直方向 1080 列的光点矩阵组成，算一下 1920x1080，也就是 207 万个光点组成。

这些光点会被记录成一个二维数组，就是一张图。每一个光点，专业上称为像素，是由红绿蓝三种基本颜色调和而成。而这三个颜色，都是以 0-255 之间的一个数字表示。也就是一个像素，就是 3 个数字，例如一个红色的像素，就是（255，0，0）。那么，要生成一张红色的图，在 1080p 的分辨率下，就需要提供 207 万个（255，0，0）红色像素的数组。而且这只是一张静态的图像，如果是视频，每秒有 60 帧图像，那一秒就需要处理上亿个像素了。

其实，在 GPU 处理图像，特别是 3D 图像的时候，倒不是一个像素一个像素处理的，而是把 3D 图形转换成可以在 2D 屏幕上展现出来的，由顶点构成的无数个三角形。然后，根据每个三角形的三个顶点，把这个三角形所覆盖区域换算成像素，然后再做颜色效果，基本上就得到了屏幕上的最终效果。

下面是用 GPU 处理一个 3D 桌子图像的绘制示意图，你大概可以理解这个处理过程。

另外我放了一个 OpenGL 做渲染的处理流程图。OpenGL 是一个渲染图形的跨语言、跨平台的应用程序编程接口 API。所谓的渲染就是绘制图形的主要操作。



从我上面说的图像处理的原理，可以看到，绘图任务需要并行处理海量数据，这对擅长做串行数据处理的 CPU 来说，既不合适，又负担很重。于是在 1980-1990 年代，图形加速卡这种外设开始出现。当 1993 年黄仁勋创建英伟达的时候，市面上有 100 多家图形加速相关的公司。这是一个可怕的在性能、标准、市场层面的混战，当时的市场主要是游戏和 PC 市场。

当然 PC 和游戏市场的残酷，也给了英伟达两个非常宝贵的经验：第一，超摩尔速度，就是被业界称为“黄氏定律”的：“半年更新，一年换代”；第二，就是软硬件之间的标准的重要性。

在游戏市场，英伟达和竞争对手竞争的早期，微软的 Direct X 可以说是能左右游戏开发生死的翻云覆雨手，它是一套优秀的应用程序编程接口，既为软件开发者提供与硬件的无关性，又为硬件开发提供策略。说白了，它就在做软硬件之间的标准。

这也是英伟达花 10 年苦功，投入 CUDA 软件生态建设，把软硬件之间的标准，移到了自己一侧的动力。关于 CUDA，我下面会讲到。

对于 GPU 开局这部分内容，当然我的重点不是分析传统的 GPU 市场，老实说，回顾英伟达的公司历史，至少从股价市值的方面，市场给予了 GPU 这个产品正常认可度。在 2016 年之前，英伟达的市值和跟营收曲线，基本上相符。**GPU 这个产品存在的价值，就是从 CPU 上卸载图形处理工作，是依附在 PC 和游戏机这两个市场的。因此长期以来，英伟达是被放在 Intel 这个坐标系下评价的。这也是为什么如果固守 PC 和游戏机这两个已经成熟的市场，英伟达的发展有限。**

英伟达的转折点，或者用理论一点的术语描述，英伟达的第二曲线是我下面要说的 GPGPU 与 CUDA。

GPGPU：点亮并行计算的科技树

2007 年，英伟达首席科学家 David Kirk 非常前瞻性地提出 GPGPU 的概念，把英伟达和 GPU 从单纯图形计算拓展为通用计算，强调并行计算，鼓励开发者用 GPU 做计算，而不是局限在图形加速这个传统的领域。GPGPU，前面这个 GP，就是 General Purpose 通用的意思。

而且英伟达这样一个芯片公司，以 12 分的努力和投入，开始建设 CUDA 这样的软件框架。GPGPU 和 CUDA 让英伟达从计算机图形学，开始走出来，把并行计算这个重要任务记在自己的名下。

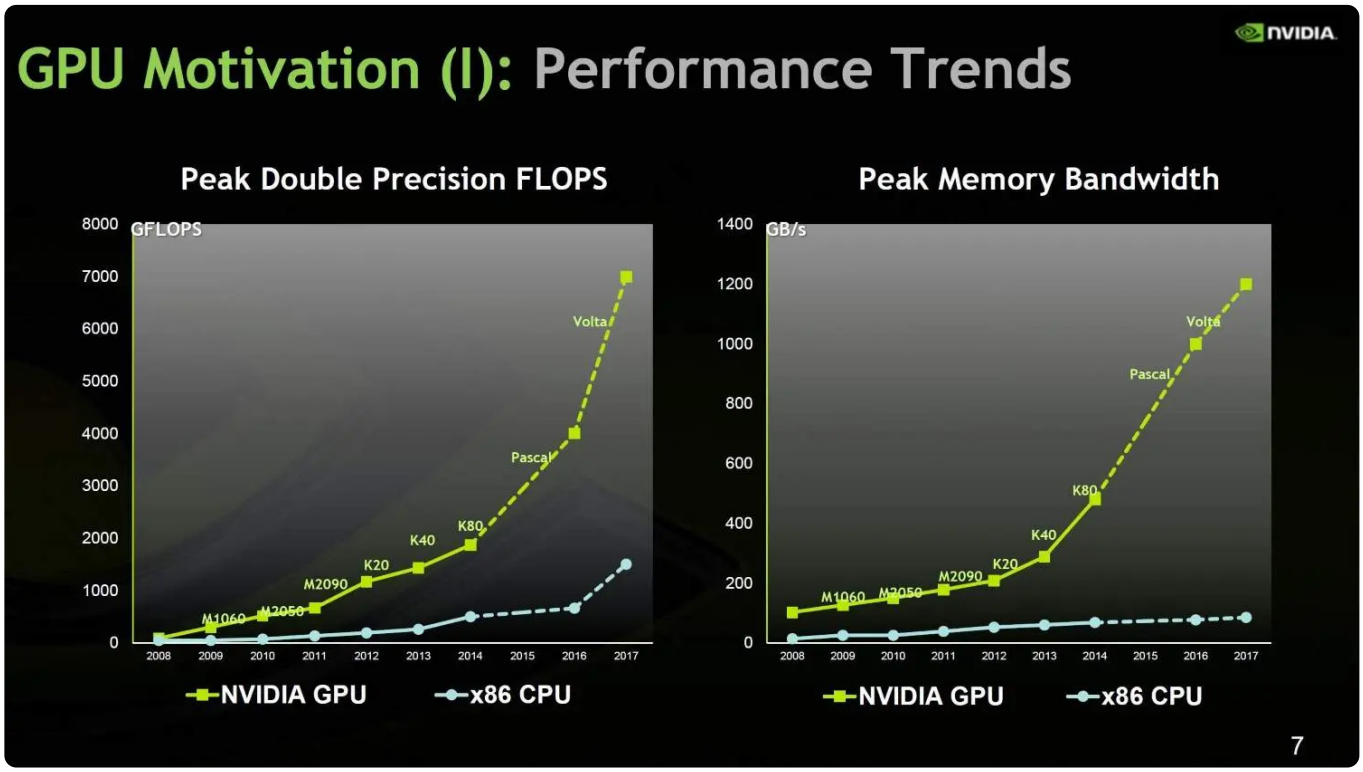
这里我要重点说一下 CUDA（Compute Unified Device Architecture，统一计算架构），CUDA 不仅仅是一个 GPU 计算的框架，它对下抽象了所有的英伟达出品的 GPU，对上构建了一个通用的编程框架，**它实质上制定了一个 GPU 和上层软件之间的接口标准**。

前面有提到，在 GPU 市场的早期竞争中，英伟达认识到软硬件之间的标准的重要性，花了 10 年苦功，投入 CUDA 软件生态建设，把软硬件之间的标准，变成自己的核心竞争力。英伟达可以说是硬件公司中软件做得最好的。

同样是生态强大，Wintel 的生态是微软帮忙建的，ARM-Android 的生态是 Google 建的，而 GPU-CUDA 的生态是英伟达自建的。

这个标准有多重要？这么说吧，一流企业定标准，二流企业做品牌，三流企业做产品。**在所有的半导体公司中，制定出软件与硬件之间的标准，而且现在还算成功的，只有 3 个，一个是 x86 指令集，一个是 ARM 指令集，还有一个就是 CUDA 了。**

2007 年，英伟达推出了基于 CUDA 的 GPGPU beta 版，之后公司的所有 GPU 都支持 CUDA，因此其实 GPGPU 和 GPU 可以完全通用。各种编程语言的工程师纷纷用英伟达的 GPU 进行开发，大大增强了 GPU 的开放性和通用性。而且因为英伟达是用 GPU 进行并行计算这个领域的第一，因此所有这个领域的软件、应用，都支持了 CUDA，CUDA 实际上成为并行计算的事实标准。



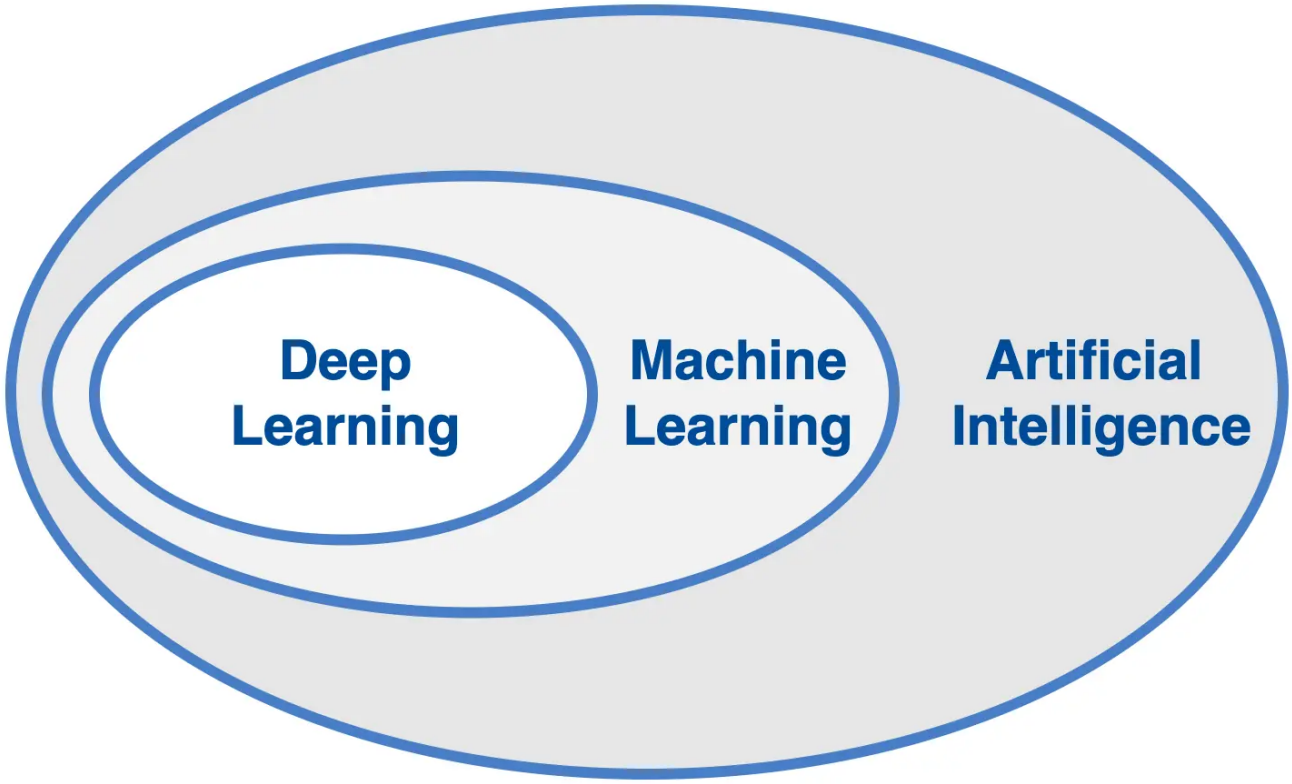
GPGPU 提出之后，在提供算力这个方面上，每个英伟达的 GPU 都有了一个统一的框架。英伟达在算力和存储带宽两个方向上，对比 CPU，都以超 10 倍，甚至 100 倍的优势领先，如上图。GPU 相对 CPU 的 TOPS per Watt（花费每瓦特电能可以获得的算力）的差异竞争优势，它的本质就是将晶体管花在计算上，而不是逻辑判断上。

在提出 GPGPU 和做 CUDA 的时候，英伟达面向的科学计算，就是冲着 HPC 和超算这个市场去的。也确实挺成功的。2020 年超级计算机 TOP500 更新榜单，可以看到 TOP10 的超级计算机中有 8 台采用了英伟达 GPU、InfiniBand 网络技术，或同时采用了两种技术。TOP500 榜单中，有 333 套（三分之二）采用了英伟达的技术。

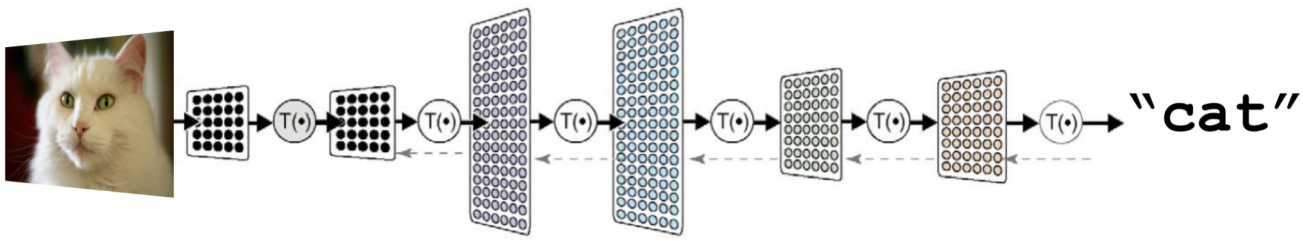
但是 HPC 和超算市场有限。有易用统一的 CUDA 编程标准，有算力和存储带宽都远超 CPU 的计算能力，有明确的技术优势，万事具备，英伟达的腾飞就差一个应用面更广大的通用问题了。

AI 的风口下，英伟达腾飞

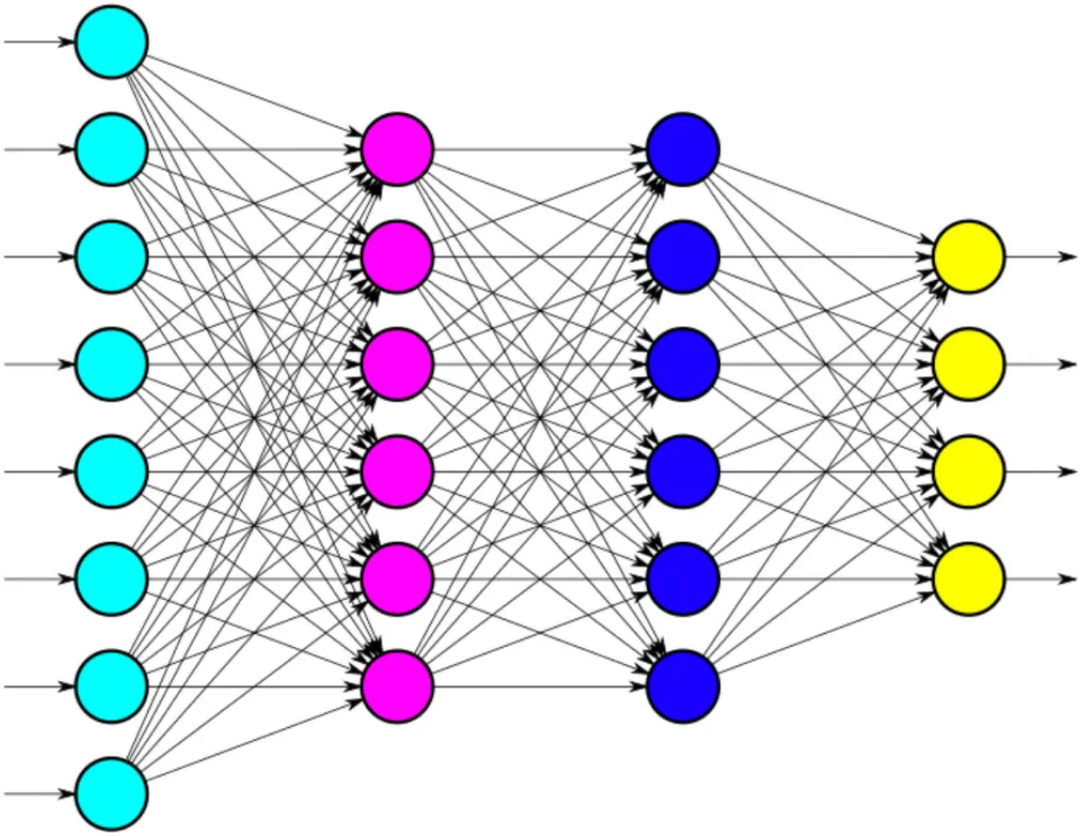
其实，**悬而未解的重要通用问题**是一直存在，这就是人工智能。如果说计算机科学有几大前沿方向，人工智能一直排在前列。关于人工智能，整个计算机行业的人，有过几轮的努力，很多不同的研究方向，其中一个方向是机器学习，而深度学习就是机器学习的一个最重要最成功的方式。下面一张图，把人工智能、机器学习和深度学习的关系表明了。



深度学习实现人工智能，就是拿神经网络模拟大脑，把推理判断过程转化成简单的，可以“训练”的数学方程。我们就以下面的从照片中识别“猫”的这个案例解释一下：



首先，准备一个充满数学公式的深度神经网络，如下图，然后提供大量猫的图片给这个网络做训练，这个网络通过层层计算被训练识别这些图片中哪些是猫，哪些不是。当这个网络训练好了，再拿一张新图片过来，这个网络就能推理出来，这张图片是不是猫。



深度学习依赖于数学和统计学计算。深度学习中的用到的各种神经网络，例如人工神经网络（ANN）、卷积神经网络（CNN）和循环神经网络（RNN），都是典型的并行结构，每个节点的计算简单且独立，但是数据庞大，通常深度学习的模型需要几百亿甚至几万亿的矩阵运算。因此深度学习的成功有两个要素：一，海量的训练数据；二，高速的并行计算能力。

几百亿甚至几万亿的矩阵运算，这一听，就很 GPU，因为 GPU 在进行图像渲染时，就是做大量的矩阵乘法运算。

每秒进行大量的矩阵乘法运算，这就是图形渲染和深度学习统一的地方。

统一的问题，给出统一的答案：英伟达 GPU。

搞算法的科学家，在用 CPU 的年代里无所作为，于是努力改进自己的算法，希望能少用一点计算，但并没有太大效果。后来，有人用上了 GPU，算法不变，算力提升 10 倍，效果立马就不一样了。混沌大学创办人李善友总是说，要注意差 10 倍的那种变量，一个方向上出现了差 10 倍的一个变量，一定有大事要发生了！

我非常赞成中国著名 AI 专家、格灵深瞳公司 CEO 赵勇博士的评价：有人说是深度学习成全了英伟达的 GPU，其实我认为，反而是 GPU，成全了深度学习。

赵勇博士讲得很清楚了，我直接把原话放在这里：“如果没有英伟达的 CUDA 平台，科学界证实深度学习巨大潜力的时间不知道还要推迟多久。更难能可贵的是，GPGPU 技术使得在 PC 级别的计算机上进行高密度的高性能运算成本大幅降低，以至于一个普通科研人员的台式电脑都有可能部署上万个并行处理内核。这使得深度学习技术迅速地在科技界发展和普及起来。可以这么说，如果没有 GPGPU，坚持研究了多年的神经网络算法的欣顿教授们，恐怕还得继续在学术界沉默不少年。”

之前我们提到过硬件行业的 3 次风口：第一次 PC；第二次手机与移动互联网；第三次就是 AI+5G+IoT。

英伟达在整个 PC 时代里它是配角，也尝试过手机，最后黯然离场。但是 AI 的大风刮来，英伟达成为人工智能这场大戏的主角英雄。为什么它能成为主角？因为它解决了 CPU 没有解决好的两个问题：并行计算、高数据吞吐能力。而且它还给自己建造了一个护城河，并行计算的软硬件之间的事实标准 CUDA。

后来的所有 AI 芯片，各种 xPU 们，在落地之前，做的第一件事，就是匹配 CUDA。而每一个流行起来的学习算法，框架，底层也全部支持了 CUDA。AI 的新算法，新框架越活跃，参与 AI 的开发者越多，CUDA 的优势就越突出。

英伟达一直保持相对于 CPU 的 10 倍的计算力，这个 10 倍速变量，让它成为一个合适问题的关键答案，也完美地验证了超越摩尔定律的都是好公司的业界通识。

当然，作为从业人士，我其实是想提一个小悖论的。GPU 的计算性能比 CPU 高 10 倍，如果人类社会的任务比例是逻辑计算，就是适合 CPU，那些任务是 1，数据计算是 10 的话，CPU 和 GPU 的比例，是 1:1，英伟达的最高点就是跟 Intel 同规模。只有逻辑计算是

1，数据计算是 100，CPU 和 GPU 的比例，才是逆转的 1:10，英伟达才会有 Intel 的 10 倍规模。你可以理解我说的么？欢迎给我留言，我们一起讨论。

总结

最后，我来给你总结一下英伟达这家公司。同样有几个要点：

1. 现在市值超越 Intel 的英伟达，是从做图形加速的 GPU 开始的，早期 GPU 一直作为 CPU 的加速器存在。
2. 但是英伟达用 CUDA，建立了并行计算的事实标准，无论是科学计算，还是计算机图形科学、计算机辅助设计，CUDA 都向上构建了最底层的编程框架，向下抽象了英伟达的所有 GPU 微架构。自此，英伟达从一个做产品的公司，进阶为做标准的公司。
3. GPU 和机器学习，相互成就，替人类打开了人工智能方向的一扇全新的大门，人类世界对英伟达的回报，就是超出业界预计的高市值。

这就是英伟达，一个半导体行业的老兵，三步封神之路。

思考题

你觉得英伟达的护城河深么？这波 AI 的大风，能把英伟达送到多高？那些号称解决 AI 问题的 xPU，有优势么？欢迎在评论区给我留言。

分享给需要的人，Ta订阅后你可得 20 元现金奖励

赞 5 提建议

© 版权归极客邦科技所有，未经许可不得传播售卖。 页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 08 | 台积电：摩尔定律的忠实执行者

下一篇 10 | 德州仪器：模拟芯片的无冕之王

更多学习推荐

精选留言 (6)

写留言

openlinuxer
2021-06-07

异构计算是群雄逐鹿，gpu,npu,fpga根据用户需求（性能及能耗）都有机会，但市场占有率还属gpu，其他两个只能做云，要不做边，因为除了加速芯片还得考虑框架，即生态，毕竟gpu有cuda加持，npu,tpu,fpga类似的框架及驱动相关生态还差得远，追上英伟达之路一厘不少，但值得期待。

展开

3

Geek_ae1337
2021-06-07

哇，终于说到皮衣教主了哈！老师提出有关未来需要更多的逻辑计算和数据计算之间的比例悖论很有意思。我感觉这个数字会大于10，甚至大于100；其中一个主要原因是逻辑计算的摩尔定律已经减缓了，每秒每瓦算力的成本下降速度减慢甚至停滞会拖累未来需求，但是数据计算的摩尔定律要仁慈的多，反而会促进需求。所以现在的AI XPU们多在数据计算上用功。这也是我建议英特尔在GPU上下功夫的原因。翻一下历史，好像英特尔90年...

展开

作者回复: Ampere是做arm server CPU的。
我个人非常看好Graphcore

1 1

分清云淡
2021-06-07

邵博士讲得真好，不过这波市值应该有80%是靠挖矿撑起来的，当然长远来看还是得靠人工智能和深度学习

作者回复: 应该不至于。我感觉投资界对挖矿，其实评价不高。

2 1

2021-06-11

英伟达的护城河很深，CUDA和GPU框架相辅相成。英伟达不仅设计芯片，又做出了统一的框架标准，基于这个标准化的架构，生态可以繁荣发展。
XPU的优势在于算力高，目前的移动处理器都倾向于在芯片架构里加入XPU单元，来扩展AI算力。缺点在于需要芯片厂家搭建软件架构，生态伙伴也需要不停适配各家的芯片，这种模式难以做大市场规模。

展开

作者回复: xPU的生态建设，需要时间，没办法。

Geek_5b9675
2021-06-08

如果我们同意x86不是通用计算的唯一答案，那么GPU大概也不是机器学习的唯一答案。

作者回复: 这个角度很赞。唯一，确实达不到唯一，但是会不会一直保持80%以上市场份额的优势呢？

水天一线
2021-06-07

Google的NPU号称是自己设计的架构和硬件，Alphago是一个典型的应用。目前图像识别和语音识别是有实际产品面世，从自动驾驶到智能问答，这些主要面对的应用方面，而对背后提供计算能力的公司却鲜有提及，英伟达制作芯片，那么类似于传统PC厂商，提供

硬件支持是否会诞生一批新的企业，比如制造矿机的公司，很有可能下一个这样的机会再这些矿机公司。

展开

作者回复: 感觉还是老牌公司，在自动驾驶芯片上在发力，中国的startup还多一点，美国看着不算多。